DOCUMENT RESUME

ED 156 141 IR 006 074

AUTHOR Klingbiel, Paul H.
TITLE Machine-Aided Indexing in Practice: An Encounter with Automatic Indexing of the Third Kind.
INSTITUTION Defense Documentation Center, Alexandria, Va.
PUB DATE 77
NOTE 11p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Automatic Indexing; *Data Bases; Documentation; *Federal Government; History; Information Science; Management; Research; Technical Reports
IDENTIFIERS *Defense Documentation Center VA; United States

ABSTRACT

This three-part report includes a brief history of the Defense Documentation Center (DDC) with a description of the collections and their accessibility; categorization of automatic indexing into three kinds with a brief description of the DDC system of machine-aided indexing; and an indication of some operational experiences with the system. (Author/MBR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED156141

IR

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

MACHINE-AIDED INDEXING IN PRACTICE:

AN ENCOUNTER WITH AUTOMATIC INDEXING OF THE THIRD KIND

Paul H. Klingbiel

Defense Documentation Center
Cameron Station
Alexandria, Virginia 22314

IR000074

INTRODUCTION

The presentation falls into three distinct parts. First, a brief history of DDC, a description of the DDC collections, and a statement of collection accessibility; second, a categorization of automatic indexing into three kinds with a brief description of the DDC system of Machine-Aided Indexing; and third an indication of some operational experiences with the system.

PART I

- A Brief History of DDC.

The origin of the Defense Documentation Center can be traced back to 1926 when a Technical News Service was started by two Air Corps Engineering employees at McCook Field in Dayton, Ohio. However, the more usual origin is placed in December 1945 when the Air Documents Division of the Intelligence Department of the Headquarters, Air Technical Service, Wright Field, Dayton, Ohio, took over some 800,000 European documents, most of which were "captured" during World War II. Captured Japanese documents were added. In May 1951 the organization was renamed the Armed Services Technical Information Agency (ASTIA). Air Force support was removed in July 1963 when the organization became a field office of the Defense Supply Agency (now Defense Logistics Agency). Simultaneously the name was changed to the Defense Documentation Center.

- The DDC Collections.

Through 1960 DDC's collection was a document collection. Since then, DDC has added three nondocument data bases.

The salient features of the DDC data bases, the text stored in machineable form, the retrieval files, and the vocabulary (as of 31 Dec 77) are presented in historical order.

1. The Technical Report (AD) collection (the original DDC data base).
   - 875,000 reports under automated bibliographic control
   - 3,000,000 words of text in machine readable form annually
   - 178,000 subject access points with 10,500,000 postings
2. The R&T Work Unit Information System.
   - Technically-oriented summaries describing research and technology projects currently in progress at the work unit level
   - 115,500 summaries under automated bibliographic control
   - 1,725,000 words of text in machine readable form annually
   - 182,700 subject access points with 2,000,000 postings

3. The R&D Program Planning Information System.

- Planning documentation in summary form at the Project and Task level
- 27,800 summaries under automated bibliographic control
- 1,100,000 words of text in machine readable form annually
- 13,450 subject access points with 518,000 postings

4. Independent Research and Development.

- Research information directed toward continual advancement of technical competence performed by industrial organizations and which is not wholly funded by a Defense Department agreement (the data is considered highly proprietary)
- 18,900 summaries under automated bibliographic control
- 1,800,000 words of text in machine readable form annually
- 41,600 subject access points with 504,000 postings

In summary

- 1,037,200 "accessions" under automated bibliographic control
- 7,625,000 words of text stored annually in machine readable form
- 309,600 subject access points with 13,522,000 postings

  NOTE: The arithmetic sum of the subject access points as listed above is 415,750. This counts duplicates.

- New "accessions" are added at the rate of 49,500 annually

Certainly, local organizations have collections equal to or larger than those of DDC. Full-text systems store many more words of text annually. The New York Times data base provides 700,000 access points. Nevertheless, DDC has already automatically indexed 20,000,000 words of text, and with current accession rates, will have indexed 100,000,000 words of text at the end of 10 more years.

- Collection Accessibility. (1)

Research and development activities within the United States Government and their associated contractors, subcontractors, and grantees, with current Government contracts, are eligible to receive most of the information from the DDC data banks. In addition, research and development organizations without current contracts may become eligible for service by a military service authorization under the Defense potential contractors programs. There are collections, however, which contain proprietary data or information compiled for the specific purpose of DoD management decisions which are made available to Defense components only.

DDC serves the general public indirectly where unclassified and unlimited documents are concerned, through a special arrangement with NTIS, Department of Commerce. More than half of the technical reports received in DDC are made available through this special arrangement. These reports are included in the NTIS publication, Government Reports Announcements.

PART II

- Automatic Indexing.

Of course, there is something ambiguous about "automatic indexing" or "machine-aided indexing" or any similar phrase. It is that ambiguity that the title of this paper purports to clarify. I am suggesting that there are at least three different kinds of automatic indexing. I shall discuss them in order.

AUTOMATIC INDEXING OF THE FIRST KIND.

Automatic indexing of the first kind is to display selected items in an alphabetical array together with a locator, page, line, chapter, etc. for a body of written material: titles, abstracts, books.

KWIC indexing of titles is automatic indexing of the first kind. Nothing is extracted as a surrogate. All words of the titles to be indexed via the KWIC method are categorized either as Display or No Display. The titles are then arranged in alphabetical order of the Display words.

Most so-called full-text indexing is also indexing of the first kind. All but a few words of the text are associated with locations. No words or phrases are extracted from the text to act as a surrogate of the text.

AUTOMATIC INDEXING OF THE SECOND KIND.

This method of automatic indexing goes beyond the display and locator concepts. Words and phrases are extracted from text to act as surrogates. If there is a display, it is an alphabetical display of the extracted items only and usually does not include a locator. Of course, if the extracted terms are placed on an inverted file, then a locator is present. The prime distinction between indexing of the first and second kinds is that in the latter case, surrogates are extracted and stored independently of the text, and the text can be searched only by the surrogates.

The process of text extraction may be statistical, linguistic, or a combination of both. Whatever the technique, natural language words and phrases are selected and are used without any further modification. Such devices as word truncation are used in the retrieval process and not as an essential part of the indexing process.

AUTOMATIC INDEXING OF THE THIRD KIND.

Automatic indexing of the third kind involves the automatic extraction of words and phrases from text, and the nontrivial automatic transformation of those natural words and phrases into a finite, controlled vocabulary of words and phrases. The end result of the process is the assignment of index terms to a text as a surrogate where the index terms do not necessarily match any of the words of the original text.

Such nontrivial automatic transformation implies the capability of extracting from text a phrase like "strategic balance" and automatically assigning as index terms the two controlled vocabulary phrases "Military strategy" and "Balance of Power"; or the automatic transformation of the natural language phrase "Animal Hosts" to "Hosts(Biology)".

MACHINE-AIDED INDEXING AT DDC

The discussion is keyed to the figure which illustrates the principal components of the operational system.

Text, in machine readable form, is read by the MAPS program one word at a time in normal text word order. As each word is read, the word is located in the Recognition Dictionary. If the word is found in the Dictionary, an instruction is obtained as to the further disposition of that word: a word may be held or rejected. If the word is held, the next word is read and looked up in the Dictionary. This process continues, one word at a time until either the word under consideration can not be found in the dictionary or the word is found and labelled as a reject.

A reject stops the reading of words from text and stops the phrase selection process. When stopped, a storage location within the program is checked to see if a phrase has been selected. If not, the reading process begins again.

When a phrase has been identified, its format (syntactic structure) is matched against a dictionary of formats held in the MAPS program. If there is a match the natural language phrase is passed on to the Natural Language Data Base. If there is no match the phrase is printed out on an error list.

The Natural Language Data Base is a thesaurus held in mass storage and used as a translator between the natural language phrases extracted from text and the controlled vocabulary used for indexing and retrieval. Index/Retrieval terms are assigned to accessions by the NLDB whenever an incoming natural phrase matches an entry in the NLDB. Depending upon the data base, as few as 55 percent or as many as 95 percent of the incoming phrases match the 250,000 entries in the NLDB. Phrases that do not match are printed out for review as possible additions to the NLDB. Based on the growth curve to date, the NLDB will grow to one million phrases before it sensibly stabilizes.

The controlled posting terms assigned by the NLDB to each accession are manually Post Reviewed accession by accession for either deletion of superfluous or misleading terms or the addition of terms not automatically assigned.

In summary, machine-aided indexing as practiced at DDC consists of four distinct steps: word categorization, phrase selection, posting term assignment, and post review. As such it is a primitive operational version of Automatic Indexing of the third kind.

PART III

- Practical and Management Consideration of Automatic Indexing of the Third Kind.

1. Work loads.

Can the operational system keep up with the work loads? All of the text normally stored by DDC annually in its four data bases can be sent through the system in 26 hours of CPU time. Wall time is obviously much greater, but is still small in comparison to the time required to manually index the same material. The work load to maintain the Recognition and Format Dictionaries utilized by the MAPS program is less than 1 man-year per year. The work load to maintain the NLDB must be managed so as not to exceed resources. Automating this last task, which is underway, will lift DDC from a primitive to a full blown version of Automatic Indexing of the Third Kind.

2. Cost.

Certain costs associated with indexing are borne by DDC whether the indexing is accomplished manually or automatically. For instance, for the last 20 years, DDC has maintained a lexicography staff of about six individuals to build thesauri and act as a clearing house for new terminology. That cost is maintained under automatic indexing. The text that is stored in the computer and used by the MAPS program to extract phrases would be keyboarded and stored even if DDC were to index all of its material manually.

One measure of cost of automatic indexing is to compare the number of personnel required to manually index the four DDC data bases contrasted with the costs unique to automatic indexing. An estimate can be obtained based on experience with the Work Unit (DD 1498 file). Manual indexing required about 3 man-years of effort. Post Review of automatic indexing should require about 1.5 man-years of effort.

It seems reasonable that the direct indexing cost is about 50 percent of manual indexing (in salaries) plus the additional computer cost to run MAPS and the NLDB. If the latter cost does not exceed the 50 percent salary saving, then there is a net dollar saving.

However, there are other considerations of an intangible nature in terms of cost. One of them is the operating level of the system. This is addressed under the next topic.

3. Recall/Precision Values.

Every manager would like a fix on the level of performance of his system. There is no measure of performance that is universally accepted. Recall and precision figures are subjective to some degree no matter how derived. In the case of automatic indexing one is attempting to judge whether retrieval is positively or negatively affected when compared with manual indexing. A before and after comparison is implied.

DDC has seldom performed Cleverdon/Lancaster type evaluations. Consequently, DDC has no firm fix on its Recall/Precision figures over the years of strictly manual indexing. We performed a small evaluation comparison which included automatically indexing material that had previously been manually indexed. We also studied the effect of automatic indexing with and without post review. This evaluation has been published. (2) Two conclusions will be repeated here.

a. Unedited machine indexing can be comparable to manual indexing in recall and is at least equal to manual indexing in precision.

b. Edited machine indexing with a suitable search strategy is at worst equal to manual indexing in recall and always performs materially better in precision.

Basically, we concluded that post edited automatic indexing was superior both to unedited automatic indexing and to manual indexing. That post edited automatic indexing is superior to unedited automatic indexing is not surprising. It may be surprising that post edited automatic indexing is also superior to manual indexing. I believe the explanation is that a list of index terms (whether derived automatically or not) presented to a reviewer acts as a prompting device. The post reviewer is less apt to overlook a significant concept than when indexing without a list. This raises recall. Since the machine selects virtually all noun phrases and converts them to posting terms, unedited automatic indexing tends to add irrelevant terms to an accession. By removing these from the list, the post reviewer increases precision. What may be surprising is that automatic indexing even with the irrelevancies present, can be comparable to manual indexing in recall and many times is better in precision. Some of the irrelevancies apparently do not cause as much havoc as has been supposed.

4. Personnel Reactions to Automatic Indexing.

The initial reaction at DDC was very negative. Basically, the implementation of automatic indexing was seen as either ill-advised (translation - indexing is an intellectual task which can not be successfully automated) or as a threat (translation - if this succeeds my professional status and job security are jeopardized). Such attitudes may be normal when automatic indexing of the third kind is involved. The attitude at DDC was certainly more severe than necessary because of the circumstances under which the system was developed and implemented.

We have just gone through a series of intensive, almost one-on-one training sessions to inform the analysts (both indexers and retrievers) of the strengths and weaknesses of the system, the details (components) of the system, and the points of human interface with the system. Attitudes have changed from negative to (at worst) neutral and in some instances to interest and enthusiasm. Psychologically, the task is to get analysts to view automatic indexing as the result of human instruction to the computer, seen as a tool, rather than to see automatic indexing as a competitive product of a mechanical adversary. With such an attitude change, automatic indexing of the third kind becomes an intellectual challenge: What is the limit to which an inanimate, nonsentient tool can be pushed in the performance of a task which has the surface appearance of requiring human judgment?

5. Skill Mix.

No particular formal education or training is required. A component of MAPS contains a right linear phrase structure grammar with 61 rules. The grammar was written under my direction by a professional librarian who had no formal linguistic training. The grammar was programmed for our computer by an individual who had no prior experience with formal grammars.

The basic operational requirement is hands-on-training to a point where the output can be traced back through the system to its origin and subsequent synthesis. Such insight prevents gross errors and is the source for system improvement. The skill requirements for development require linquistic, mathematical, programming, and information science components.

6. Potential Outside of DDC.

At the moment all of the programs are written in languages associated with UNIVAC equipment. Other agencies have expressed interest in using the DDC system, but have been prevented because of programming language incompatibility. The system written in a higher order language would be very versatile and could be used, with very minor modification, to index material quite different than that in the DDC collection. Substantial changes might be required if lengthy text (full documents, books, etc.) were to be indexed, depending upon the purpose the index was to serve.

BIBLIOGRAPHY

1. Federal Scientific and Technical Communication Activities, Oct 1977, PB 272 800, P. 23. (Available from NTIS)

2. Klingbiel, P.H., Rinker, C.C., Evaluation of Machine-Aided Indexing. Information Processing and Management Journal, Vol. 12, No. 6, 1976.